Available online www.jocpr.com

Journal of Chemical and Pharmaceutical Research, 2014, 6(5):352-359

Research Article

ISSN : 0975-7384
CODEN(USA) : JCPRC5

Research of liver cancer detection based on improved K-NN algorithm

Jianhua Liu[1], Jianwei Wang[1] and Wenjuan Bu[2]

[1]*Software School of North China University of Water Resources and Electric Power, Zhengzhou, Henan, China*
[2]*Information Engineering University, Zhengzhou, Henan, China*

ABSTRACT

As a simple and effective classification algorithm, k-Nearest Neighbor (K-NN) algorithm is widely used in many fields. In order to improve the efficiency of classification in Liver Cancer Detection, the Principal Component Analysis (PCA) method is applied to the K-NN algorithm, which selects the effective features efficiently. Building new attribute sets and applying new effective features to K-NN classification separately, to obtain the correct classification rates of new effective features. Then the correct classification rates are applied to the K-NN Algorithm for classification as the distance weight. The improved K-NN Algorithm has been applied to Liver Cancer Detection, and the experiment indicated has obtained the good effect.

Key words: K-NN; Principal Component Analysis; Distance Weight; Liver Cancer Detection

---

INTRODUCTION

Distinguishing normal liver and liver lesions, then determining the type of liver disease is the final problem in computer aided diagnosis system on liver cancer. At present, to solve this problem, the principal way is using texture feature and shape feature of liver, combine with classification algorithm of data mining[1]. Being a classical classification algorithm, the K-NN algorithm is applied widely. In this paper, based on the further analysis of the traditional K-NN algorithm, the distance model of K-NN algorithm was improved, and then using the improved K-NN algorithm into the process of classification and identification of liver cancer. Experimental results indicate that this new improved K-NN algorithm has significant effectiveness.

ANALYSIS OF TYPICAL K-NN ALGORITHM

K-NN algorithm is a comparatively mature theory method. The main classification principle is as follows: In the feature space, if most of the samples in k-Nearest Neighbors of the sample to be classified belong to a category, the determination of the sample to be classified is that this sample also belongs to the category [2].

The fundamental ideas of Classification algorithm of K-NN is as follows: For a given unknown sample, the calculation of distance between training samples and unknown sample is proceed at first, then select k-Nearest samples from training samples set with the nearest distance between training sample and the given sample. And the classification of the unknown samples is according to these k-Nearest samples' category.

The algorithm assumes that all the samples corresponding to the points in the n-dimensional space Rn. In the n-dimensional space, the feature vector set for sample x is as follows:

$<a_1(x),a_2(x)\ldots a_n(x)>$ (1)

Notes: an(x): the nth property values of sample x.

The distance between the two samples x_i and x_j is indicated as $d(x_i, x_j)$, Expressed by Euclidean distance is as follows: $d(x_i, x_j) = \sqrt{\sum_{r=1}^{n} \left(a_r(x_i) - a_r(x_j)\right)^2}$ (2)

According to the formula (2), the k-nearest neighbor of the sample to be classified is selected. The key points play an important role, in the process of application of K-NN algorithm model are as follows:

(1) The selection of training samples;

(2) The mathematical model for calculation of distance;

(3) The selection of the value of K;

(4) The basis for classification;

These key points above have important influence on the accuracy of classification of K-NN algorithm. These influencing factors reflect in the progress of the application of classification are as follows:

(1) The features with no effect on the classification or with little effect on the classification are involved in the progress of the application of classification, which reduce the accuracy of classification;

(2) The features involved in the progress of classification invariably with different influence degree. These features will have adverse effects on the classification, on condition that different weight values are not attached to these features according to these influence degree;

(3) The different distance models, reflecting relationships among features, have different degree of influence in the progress of classification. The selection of different distance model will obtain different accuracy of classification.

Most of the improved K-NN algorithms are according to these factors above and combining with the specific application of different fields. Removing features with little effect on the classification and improving the distance model are the hotspots of the improvements of K-NN algorithm.

THE PRINCIPLE OF IMPROVED K-NN ALGORITHM

In this paper, removing features with little effect on classification and improving the distance model is the main direction of K-NN algorithm improvement.

The basic principle of improved K-NN algorithm is as follows: Firstly, the PCA method was introduced into the progress of classification of K-NN algorithm, in order to remove these features with little effect on the classification, and obtain new features described as principal component [2, 3]; secondly, different proportion values were attached to new features according to their influence degree, and participate in the computational of distance model, in order to obtain the improved distance model, which reflects the degree of impact of each new feature in classification; finally, colligating the steps described above, choosing the appropriate value of "K" and classifying the samples.

AThe Principle of PCA

To remove features with little effect on classification, we introduced the PCA into the process of classification of K-NN algorithm. The goal of PCA is to identify the most meaningful basis to re-express a data set. The hope is that this new basis will filter out the noise and reveals hidden structure. The PCA produces a series of unrelated new comprehensive indicators, by constructing appropriate linear combinations of the original features of samples, and then selects new comprehensive indicators produced in the step described above, which keep the information of the original indicators of sample as much as possible, finally, using these selected new comprehensive indicators, instead of the original indicators of samples, to analyze and solve problems[4].

The principle of PCA is as follows:

Assume that the original features of sample (the original indicator) are described as follows: $x_1, x_2, x_3, ...x_m$, and the new comprehensive indicators, corresponding to the original indicator, are described as follows: $y_1, y_2, y_3, ...y_p$ ($p \le m$), and the relationship between the two indicators is as follows:

$$\begin{cases} y_1 = u_{11}x_1 + u_{12}x_2 + \cdots + u_{1m}x_m \\ y_2 = u_{21}x_1 + u_{22}x_2 + \cdots + u_{2m}x_m \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ y_p = u_{p1}x_1 + u_{p2}x_2 + \cdots u_{pm}x_m \end{cases} \quad (3)$$

$$U = \begin{bmatrix} u_{11} & u_{21} & \cdots & u_{p1} \\ u_{12} & u_{22} & \cdots & u_{p2} \\ \vdots & \vdots & & \vdots \\ u_{1m} & u_{2m} & \cdots & u_{pm} \end{bmatrix} \quad (4)$$

$$Y=UTX \quad (5)$$

$$Y = (y_1, y_2, y_3, \cdots y_p) \quad (6)$$

The selected new comprehensive indicators $y_1, y_2, y_3, ...y_p$ are described as the 1st, 2nd, 3rd…, Pth principal component of the original indicators $x_1, x_2, x_3, ...x_m$, and y_1 accounted for the largest proportion in the total variance, and the proportion of the rest indicators are decreasing according to the order. In the process of analysis of practical problems, the principal components with larger proportion are selected, in order to reduce the number of variables, grasp the principal contradiction, and simplify the relationship between variables.

BThe Principle of Improved Distance Model

The weight values of every principal component were also introduced into the distance model, in order to improve the effectiveness of distance calculation. The basic principle of improved distance model of K-NN algorithm is as follows:

1) Every single principal component was used in classification of classical K-NN algorithm separatelyat first (the 1st classification), and the correct rate of classification was recorded as the weight value of principal component.

2) Using the recorded weight value of corresponding principal component to participate in the classification of K-NN algorithm (the 2nd classification).

The improved distance model is as follows:

$$d(x_i, x_j) = \sqrt{\sum_{r=1}^{n} w_r (a_r(x_i) - a_r(x_j))^2} \quad (7)$$

Notes:

$a_r(x)$: The rth principal component of sample x;

w_r : The correct rate of the rth principal component of sample χ obtained in the 1st classification.

$d(x_i, x_j)$: The distance between any two samples x_i and x_j .

C The Basic Steps of The Improved K-NN Algorithm

The basic steps of the improved K-NN algorithm are as follows:

1) The features of the samples to be classified are analyzed with PCA, to obtain every components and their proportion, and determine the principal component to participate in classification of K-NN algorithm, which can decrease the number of components, and reduce the computation and computational complexity of classification.

2) Every single principal component is used in classification of classical K-NN algorithm separately, and then the correct rate of classification is recorded as weight values (w), in order to participate in the distance calculation of K-NN algorithm in the next step.

3) The distance between any two samples are calculated with the weight values (w) by the formula (7), and then choose the appropriate value of "K" by experience, in order to obtain the results of classification.

APPLICATION OF IMPROVED K-NN ALGORITHM IN CLASSIFICATION AND IDENTIFICATION OF LIVER CANCER

Texture feature and shape feature of liver is the main basis of classification and identification of liver cancer. Generally, normal liver and liver with diseases are distinguished by texture feature, and the types of liver cancer are distinguished by shape feature. In this paper, the improved K-NN algorithm is applied to classification based on texture feature and classification based on shape features respectively [5-10].

A Classification Based on Texture Features

Texture generally refers to the gray changed rules of image pixels (or sub-region) people observed. In image analysis, the digital features which describe the gray changes are defined as image texture features. Texture features are important features to reflect the macroscopic gray changes. Tissue texture can reflect human tissue normal or not. Texture pattern of normal tissue is the same. But texture pattern of diseased tissue is different. Texture ofhealthy tissue and diseased tissue are different in the degree of thickness and trend of distribution, So that image texture is an important feature of medical images.

Gray Level Co-occurrence Matrix (GLCM) is the most commonly used statistical analysis methods. It is an image texture feature analysis method based on the second-order conditional probability density function. It is the basis is to analyze basic image model and images arranged rules [11].

In order to describe the image texture features intuitively with GLCM, some parameters which can reflect the state of the matrix are exported. The parameters obtained by feature extraction are described as follows:

1)Energy: Energy is the sum of squares of each element of GLCM and a uniform measure of gray changes of image texture; itreflects the degree of uniformity and texture's level of coarseness of the image gray distribution. When the image texture is fine and it's evenly distributed, energy valueisbigger.On the contrary itis smaller.

2)Entropy: Entropy is a measure ofinformation content of image.It showsthe complexity level of image texture andreflectsheterogeneity of image texture.The GLCM's grayscale is almost zero and the entropy value is also close to zerowhen the image presents very smoothly and has no texture at all.On the other hand, pixel values of GLCM are equal, entropy tends toward the maximum value when the image is full of fine texture,when the imageshows less texture, elements values of GLCM are very uneven, and the entropy value is smaller.

3)Contrast:Contrast is also known as moment of inertia. It is moment of inertia near the main diagonal of GLCM. It measures how the values of the matrix are distributed and level of local changes within image, reflects image clarity and groove depth of texture. The deeper texture grooves are, the greater contrast is, the more clarity visual effect gains. On the contrast, when grooves are shallow and the visual effect is unclear, the contrast is small. Contrast ratio evaluates gray changes of image from another aspect. The more uneven the image shows, the greater differences the gray values are, the finer textures captured, the greater the contrast value provides.

4) Correlation: Correlation measures the degree of similarity of space GLCM in row or column direction. Size of the relevant values reflects the local gray correlation of image. When the values of matrix elements are equal, value of correlation is big. On the contrary, when there is a big difference between matrix elements, value of correlation is small.

5) Local stationarity: Local stationarity is also known as inverse gap. It is used to measure the local changes of image texture. When there islacks of change between different regions of image texture, value of local stationarity is big and local is very uniform. Local stationarity reflects the homogeneity of the image texture to some extent

The steps of classification based ontexture features are as follows:

1) Using the method of PCA to calculate the components and their contributions, corresponding to the texture features are as follows:

TABLE 1. PCA ANALYSIS OF TEXTURE FEATURES

Component	Y1	Y2	Y3	Y4	Y5
Contribution Rate %	25.74	21.66	24.53	19.82	8.25
Accumulation of Contribution Rate %	25.74	47.40	71.93	91.75	100.0

The accumulation of contribution of the first four components achieved 91.75%, and the fifth component contribution rate is only 8.25%. According to the data in table 1, we can conclude that the first four components can be used as the principal components.

2) Based on the step 1, every single principal component was used in classification of classical K-NN algorithm separately (the value of "K" is "7"), and then the correct rates of classification which would participate in the distance calculation of K-NN algorithm in the next step were recorded as follows:

TABLE 2.CORRECT CLASSIFICATION RATES OFEVERY PRINCIPAL COMPONENT

Principal Component	Y1	Y2	Y3	Y4
Correct rate %	65.34	42.76	52.00	34.81

3)For 1000 samples, using the principal components (Y1, Y2, Y3, Y4) as the new features, the distance between any two samples are calculated by using"(7)", with the weight values λ which was recorded in TABLE 2. The value of "K" is 7, which was determined by experience. The result of classification was as follows:

TABLE3. CORRECT RECOGNITION RATE BASED ON TEXTURE FEATURES

Total number of sample	A	B
	300	700
The number of images identified correctly	249	604
Correct Rate %	83.00	86.29

Notes:
Column A: The number of images with normal liver;
Column B: The number of images with liver cancer;

There are 300 images with normal liver, and 700 images with liver cancer. According to table 3, we can conclude that the correct rate of recognition of normal liver was 83%, and the correct rate of recognition of liver cancer was 86.29%, both of the correct rates were more than 80%, which could meet the demand of classification and identification of live cancer in a certain extent.

B Classification Based on Shape Features

The types of liver cancer are divided as cyst, hemangioma and liver cancer. These types of liver cancer can be distinguished by shape features. The shape featuresare described as follows:

1) Perimeter P of the target area boundary

Suppose B is the target area boundary. The perimeter of B refers to the length of the smallest outer boundary contour of the connected region. The number of pixels along a boundary gives a rough approximation of its length. A method based on the 8-neighborhood chain code is employed to calculate the perimeter of boundary. Using 8-neighborhood chain code to trace the outline of the target area and record the chain code value. Suppose the number of even pixels on the outer contour was M, the number of odd pixels was N, So that, the function of calculating the perimeter of connected region is as follows:

$$P = M + \sqrt{2}N \qquad (8)$$

Although the calculation is complex and low speeded, accuracy is higher than the other methods.

2) The Length and the width of the target area

(1) The lengthof the target area(Maxd): The distance between two points on the boundary which are widely separated is defined as the lengthof the target area.

The length of the target area is defined as:

$$Maxd = \max_{i,j}[D(p_i, p_j)] \qquad (9)$$

Where Maxd could be Euclidean distance, or block distance or chessboard distance, p_i and p_j are points on the boundary.

(2)The width of the target area

The straight line between the two points with longest distance divides the target area boundary into two categories. The sum of the maximum vertical distances between the points on different parts of the boundary and the straight line is defined as the width of the target area. The width of the target area is calculated as follows:

□Record two terminal vertex of the two points with longest distance,, $A(x_A, y_B)$、 $B(x_B, y_B)$.

②Link the endpoints A B to a straight line L. Divide points of the target area into two categories: from point A to point

B, the clockwise points on the target contour can be collect as one category, and the other category contains those points on counterclockwise direction.
Calculate the longest distance x1, x2, between points from those two categories to line L respectively.
Calculate the width of the target area.

Mind=x1+x2 (10)

3) Area A of the target region
There are many methods to calculate the target area of digital image. A method based on counter chain code was adopted in this paper.Get integral for x-axis along eight-neighborhood chain code contour, then the area of the target region base on 8-neighborhood chain code can be obtained. The calculation formulas are as follows:

$$A = \sum_{i=1}^{n} dx(c_j)[y_{i-1} + \frac{1}{2} dy(c_i)] \quad (11)$$

$$y_i = y_{i-1} + dy(c_i) \quad i = 1,2,\cdots,N \quad (12)$$

Note:

y_i :The vertical axis;
N: The number of chain code value;
$dx(c_i)$ and $dy(c_i)$: Theoffset of the horizontal and vertical coordinates.

4) Perimeter and area of Minimum External Rectangle (MER)
Minimum External Rectangle (MER) is the rectangle formed by the maximum diameter and the minimum diameter. Perimeter of MER is calculated as follows:

$$p_r = 2(Maxd + Mind) \quad (13)$$

Area of MER is calculated as follows:

$$A_r = Maxd \times Mind \quad (14)$$

5) Circularity
Circularity is also known as compactness. It is used to describe the circularity of regional shape. Circularity is calculated as follows:

$$C = \frac{P^2}{4\pi A} \quad (15)$$

Note: C is circularity, P is the perimeter of regional bounder, and A is the regional area. When the region is circular, C gains the minimum value 1. When the region is long and thin strip or even more complex, value of C is smaller.

When size of the target region remains the same, the regional area and the length of the border are often used to describe the shape of target. For image which its area is known, the perimeter is smaller and closer to circle. Longer diameter means rougher surface and more complex shape. For image which its area is unknown, the bigger the circularity, the longer the diameter of unit area, and the more complex the shape of region.

6) Eccentricity
Suppost the length and the width of target region are known, Eccentricity is defined as:

K=Mind/Maxd (16)

7)Rectangularity
Rectangularity of target region reflects the filling level of target area to its minimum external rectangle.

Rectangularity is calculated as follows:

$$e = \frac{A}{A_r} \quad (17)$$

Note: A is the area of the target region; Ar is the area of minimum external rectangle of the target region.

Rectangularity e of rectangular gets the maximum value 1, e of triangle rectangular gets value 1/2, e of round get value π/4, e of ellipse get value π/2. Value of rectangularity becomes smaller for slender and curved objects.

8) Edge complexity

Edge complexity refers to the ratio of the length of irregular edge images to the length of their fitting images.Edge complexity is calculated as follows:

$$R = \frac{P}{P_r} \qquad (18)$$

Note: P is the perimeter of target region; Pr is the perimeter of minimum external rectangle of the target region.

The steps of classification based on shape features are as follows:

1) Using the method of PCA to calculate the principal components and their contributions, corresponding to the shape features are as follows:

TABLE 4. PCA ANALYSIS OF SHAPE FEATURES

Component	Y1	Y2	Y3	Y4
Contribution Rate %	54.90	32.61	3.92	8.57
Accumulation of Contribution Rate %	54.90	87.51	91.43	100

The accumulation of contribution of the first two components achieved 87.51%, and the accumulation of contribution of the 3rd and 4th component are only 12.49 %. According to the data in table 4, we can conclude that the first two components can be used as the principal component.

2) Based on the step 1, the two principal components were used in classification of classical K-NN algorithm separately (the value of "K" is "5"), and then the correct rate of classification which would participate in the distance calculation of improved K-NN algorithm in next step was recorded as follows:

TABLE 5.CORRECT CLASSIFICATION RATES OF EVERY PRINCIPAL COMPONENT

Principal Component	Y1	Y2
Correct Rate %	35.90	78.00

3) For 700 samples, using the principal components (Y1, Y2) as the new features, the distance between any two samples are calculated by the formula (7), with the weight values w, which were recorded in table 5. The value of "K" is 5, which was determined by experience. The result of classification was as follows:

TABLE6.CORRECT RECOGNITION RATE BASED ON SHAPE FEATURES

Total number of sample	A	B	C
	300	140	260
The number of images denitrified correctly	246	103	227
Correct Rate	82.00%	73.57%	87.31%

Notes:
Column A: The number of images with cyst;
Column B: The number of images with hemangioma;
Column C: The number of images with liver cancer;

There were 300 images with cyst, 140images with hemangioma, and 260 images with liver cancer. According to table 6, we can conclude that the correct rate of recognition of cyst was 82%, the correct rate of recognition of hemangioma was 73.57%, the correct rate of recognition of liver cancer was 87.31%, and the correct rate of recognition of cyst and liver cancer were more than 80%.

CONCLUSION

Combined with the specific characteristics of classification feature of live cancer, in this paper, we proposed an improved classification algorithm of K-NN, and provided the detailed description of the improved algorithm. The experiment proved that the algorithm proposed in this paper can obtain better results in classification and identification of live cancer, and the results of the experiment are as follows: 1) In addition to liver hemangioma with the accuracy as 73.57 percent which is less than 80 percent, the rest of the classification accuracy are more than 80 percent, can meet the demand of classification and identification of live cancer in a certain extent; 2) This improved algorithm is suitable for the classification with a few features, especially for classification and identification of live cancer, nonethelessits time efficiency is lower than traditional K-NN algorithm, whichmay also need a little more improvements.

REFERENCES

[1] Shi Yanling, *The Research of Liver Cancer Detection Methods Basedon Medical Image[D], North China University of Water Resources and Electric Powe*r, May **2011**.
[2] Hang Jianhua, Ding Jianrui, Liu Jiafeng, ZhangYingtao. *Journal of Electronics & Information Technology*,v 35, n3, p 627- 632,**2013.**
[3] Liu Jianhua, Shi Yanling. *Image feature extraction method based on shape characteristics and its application in medical image analysis[A]. 2011 International Conference on Applied Informatics and Communication,* **2011.**
[4] Wang Zhenghong, Zhou Ling, *microcomputer information*, v 23, n 28, p235-237, **2007.**
[5] WangJuan, ZiLinlin, Yao KangZe. A Survey of Feature Selection[J]. *COMPUTER ENGINEERING &SCIENCE*.v27, n 12, p 68-71, **2005**.
[6] Yu Mei, *The Research on Techniques of Feature Extraction forHepatic CT Images and Its Application in Retrieval [D]. Southern Medical University*, May **2012**.
[7] Zhang Xuya. *Medical ImageAnalysis Based on Feature Extraction and Machine Learning[D]. Nanjing University of Posts and Telecommunications.* **2011**.
[8] Wu Shuang. *Edge Points Based Feature Extraction Method for Medical Image Classification [D].Northeastern University*. January **2008**.
[9] Liu Jianhua, Shi Yanling. *Appl. Math. Inf. Sci*.v7, n2, p.787-798,**2013.**
[10] Yan Hongwen，Liu Xinchao．*CYCS* 2005, pp.114-119, September **2005**.
[11]GaoChengcheng, HuiXiaowei. *Application of computer system*,v 19, n 6, p195-198, **2010**.